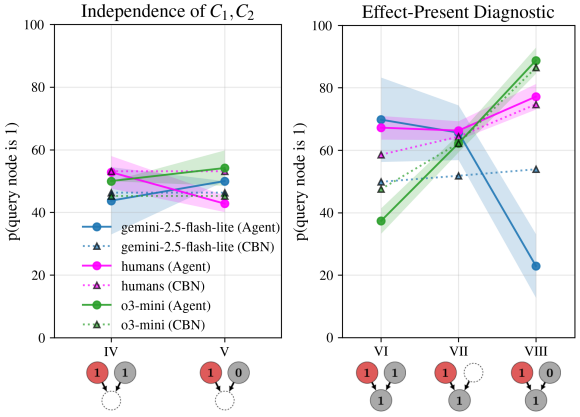
Causal Strengths and Leaky Beliefs: Interpreting LLM Reasoning via Noisy-OR Causal Bayes Nets



(a) MV *iff* $|IV - V| > 0 + \epsilon$ (b) EA iff $VIII > VI$

Figure 1: **Agents vs. CBN Predictions** Likelihood judgments that query node ● has value $1 \in \{0, 100\}$ of agents' predictions vs. their respective CBN model predictions with bootstrapped 95% confidence intervals for agents. Graphs on the x-axis visualize a subset of the conditional probability of the causal inference tasks (I-XI) where the nodes are colored according to: ● → query node that the question is asked about; ● → observed $\in \{0, 1\}$; and ◌ → no information on. Fig. 1(a) shows *Markov violations (MV)* for humans and gemini-2.5-flash-lite, as $|IV - V| > 0 + \epsilon$, visualized by non-horizontal lines, where ϵ is 0.05 in our study. o3 shows no Markov violations and perfect independence of causes. Fig. 1(b) brings about *explaining away (EA)*, iff $VIII > VI$, visualized by a postitive slope. *o3 displays perfect EA*, whereas gemini-2.5-flash-lite shows no EA and *humans show weak EA*. Experiment: Semantically meaningful (RW17) content, numeric prompt.

The nature of intelligence in both humans and machines is a long-standing question. While there is no universally accepted definition, the ability to reason causally is often regarded as a pivotal aspect of intelligence (Lake et al., 2017). Evaluating causal reasoning in LLMs and humans on the same tasks provides hence a more comprehensive understanding of their respective strengths and weaknesses. **Goals, Contributions & Methods.** Our study asks: (Q1) Are LLMs aligned with humans given the same reasoning tasks (RW17 collider tasks)? (see Dettki et al. (2025) and Rehder et al. (2017)) (Q2) Do LLMs and humans reason consistently at the task level? (Q3) Do they have distinct reasoning signatures? We answer these by evaluating 20+ LLMs on eleven $C_1 \rightarrow E \leftarrow C_2$ queries on semantically meaningful tasks (RW17) under *Numeric* (one-shot number as response = likelihood judgment of query node being one (Fig. 1)) and *Chain of Thought* (CoT; think first, then provide answer) prompting at T=0. Judgments are modeled with a leaky noisy-OR causal Bayes net (CBN) whose parameters $\theta = (b, m_1, m_2, p(C)) \in [0, 1]$ include a shared prior $p(C)$; we select the winning model via AIC between a 3-parameter symmetric causal strength (m_1=m_2) and 4-parameter asymmetric ($m_1 \neq m_2$) variant. The 3 research questions Q_i map to: human-LLM Spearman correlation ρ (Q1), task-level LOOCV-R^2 from CBN fits (Q2), parameter-signature profiling $(b, m_1, m_2, p(C))$ (Q3). This separates our work from Dettki et al. (2025) by replacing the logistic link with leaky noisy-OR, expanding the number of evaluated LLMs ($\sim 5\times$), and enabling clearer evaluations of explaining away (EA) and Markov-violation (MV) diagnostics. EA emerges in collider graphs when evidence for one cause reduces the belief in the other cause, visually represented as a positive slope in Fig. 1(b). MV occurs when the presence of one cause affects the belief in another cause, violating the independence assumption in a collider structure Fig. 1(b), visually represented by a slope e.g., humans, while `o3` shows no MV.

SOTA models establish ceiling; CoT helps others converge (Q1). Recent top-performing LLMs, e.g., `gemini-2.5-pro`, already show strong human alignment under Numeric prompting ($\rho \approx 0.85$), with only little to no improvement via CoT. Conversely, CoT significantly boosts alignment in lighter or older models (e.g., `gemini-2.5-flash-lite`: $+0.503 \rightarrow \rho = 0.845$), helping them converge to the same ceiling.

Humans are consistent reasoners; CoT improves reasoning consistency, especially for smaller & older models (Q2). On RW17, CoT yields a small but reliable lift in median reasoning consistency, raising LOOCV R^2 from 0.933 to 0.941 (+0.008, +0.91%). More importantly, CoT disproportionately helps the less consistent agents under Numeric: the lower tail rises (minimum R^2: 0.277 (gemini-2.5-flash-lite; numeric) → 0.692 (claude-3-haiku-20240307; CoT)) and the spread tightens markedly (IQR 0.116 → 0.060). Humans show high consistency across tasks, with LOOCV R^2 = 0.937 with only a narrow gap to SOTA models who achieve LOOCVR^2 of up to .99 (`gemini-2.5-pro`).

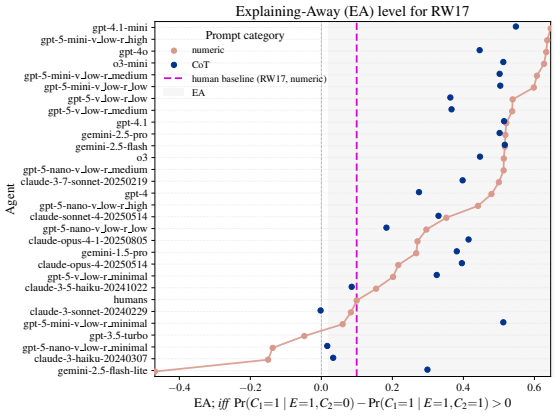


Figure 2: EA levels by agent and semantically meaningful RW17 prompts both Numeric and CoT. We don't have as many CoT responses as we have Numeric responses, hence the the gaps in the plot for CoT. The naming convention for the GPT-5 familiy is as follows: `gpt-5<suffix>_v_<verbosity-level>_r_<reasoning-effort>`.

Explaining-away is common; CoT effects are mixed (Q3). Most LLMs (27/30) show explaining-away ($EA_{raw} > 0$), and 24/30 exceed human EA levels (EA_{human}=0.09) (see Fig. 2). CoT helps agents lacking EA (e.g., `claude-3-haiku`, `gemini-2.5-flash-lite`) but can slightly reduce EA in strong ones (e.g., `gpt-4o`, `o3-mini`). A similar pattern holds for Markov violations: while eight agents show MV under Numeric, CoT improves most but can worsen others (e.g., `claude-3.5-haiku`). High-EA–no-MV agents have low leakage b (0–0.1), strong causal strength m_1, m_2 (0.75–0.99), and midrange priors, while agents with MV or weak EA show higher b (0.15–0.62) and weaker m_i (0.25–0.82).

Outlook. Next steps include extending this framework to semantically meaningless tasks and other causal structures beyond colliders to probe reasoning robustness. It should be noted that "normative" parameter regimes (low leak, strong causes) are not universally optimal and ultimately depend on the user-setting: tasks that legitimately require uncertainty about unobserved causes may warrant nonzero leak. Our prompts do not control this dimension – we neither instruct models to ignore nor to include unmentioned causes. A targeted analysis of the explanations received through CoT could provide first insights into whether and how LLMs represent and regulate them.